BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

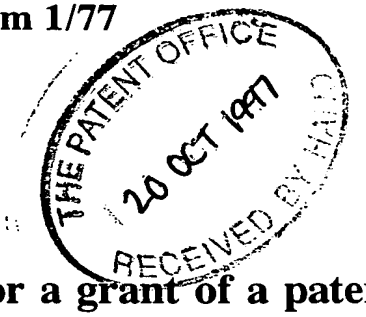
Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed P. Mahoney

Dated 12 August 2003

Patents Form 1/77

Patents Act 1977
(Rule 16)



The Patent Office

D Y & CO

210CT97 E311360-11 D02246
P01/7700 25.00 - 9722131.1

Request for a grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

1. **Your reference** — P/3038.GB ATM

2. **Patent application** *(The Patent Office will fill in this part)* — **9722131.1** — 20 OCT 1997

3. **Full name, address and postcode of the or of each applicant** *(underline all surnames)*

Medical Research Council
20 Park Crescent
London
W1N 4AL

Patents ADP number *(if you know it)* 59600700 1

If the applicant is a corporate body, give the country/state of its incorporation — United Kingdom

4. **Title of the invention** — Method

5. **Name of your agent** *(if you have one)* — D YOUNG & CO

"Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*

21 NEW FETTER LANE
LONDON
EC4A 1DA

Patents ADP number *(if you have one)* — 59006

6. **If you are declaring priority from one or more earlier patent applications, give the country and date of filing of the or each of these earlier applications and (if you know it) the or each application number**

Country	Priority application number *(if you know it)*	Date of filing *(day/month/year)*

7. **If this application is divided or otherwise derived from an earlier UK application, give the number and filing date of the earlier application**

Number of earlier application	Date of filing *(day/month/year)*

8. **Is a statement of inventorship and of right to grant of a patent required in support of this request?** *(Answer 'Yes' if:*
a) any applicant named in part 3 is not an inventor, or
b) there is an inventor who is not named as an applicant, or
c) any named applicant is a corporate body.
See note (d))

YES

9. **Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document**

Continuation sheets of this form	0
Description	30
Claims(s)	3
Abstract	1
Drawing(s)	0

10. **If you are also filing any of the following, state how many against each item.**

Priority documents	
Translations of priority documents	
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	
Request for preliminary examination and search *(Patents Form 9/77)*	
Request for substantive examination *(Patents Form 10/77)*	
Any other documents *(please specify)*	

11. **I/We request the grant of a patent on the basis of this application.**

Signature A. M.

Date 20 10 1997

D YOUNG & CO
Agents for the Applicants

12. **Name and daytime telephone number of the person to contact in the United Kingdom**

Dr A Maschio 01703 634816

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 01645 500505

b) Write your answers in capital letters using black ink or you may type them

c) If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.

d) If you answered 'Yes' Patents Form 7/77 will need to be filed.

e) Once you have filled in the form you must remember to sign and date it.

f) For details of the fee and ways to pay please contact the Patent Office.

Method

The present invention relates to methods for selecting repertoires of polypeptides using generic ligands. In particular, the invention describes a repertoire of antibody polypeptides which can be selected with a generic ligand to isolate a functional subset thereof.

Introduction

The antigen binding domain of an antibody comprises two separate regions: a heavy chain variable domain (V_H) and a light chain variable domain (V_L: which can be either V_κ or V_λ). The antigen binding site itself is formed by six polypeptide loops: three from V_H domain (H1, H2 and H3) and three from V_L domain (L1, L2 and L3). A diverse primary repertoire of V genes that encode the V_H and V_L domains is produced by the combinatorial rearrangement of gene segments. The V_H gene is produced by the recombination of three gene segments, V_H, D and J_H. In humans, there are approximately 51 functional V_H segments (Cook and Tomlinson (1995) *Immunol Today*, **16**: 237), 25 functional D segments (Corbett *et al.* (1997) *J. Mol. Biol.*, **268**: 69) and 6 functional J_H segments (Ravetch *et al.* (1981) *Cell*, **27**: 583), depending on the haplotype. The V_H segment encodes the region of the polypeptide chain which forms the first and second antigen binding loops of the V_H domain (H1 and H2), whilst the V_H, D and J_H segments combine to form the third antigen binding loop of the V_H domain (H3). The V_L gene is produced by the recombination of only two gene segments, V_L and J_L. In humans, there are approximately 40 functional V_κ segments (Schäble and Zachau (1993) *Biol. Chem. Hoppe-Seyler*, **374**: 1001), 31 functional V_λ segments (Williams *et al.* (1996) *J. Mol. Biol.*, **264**: 220; Kawasaki *et al.* (1997) *Genome Res.*, **7**: 250), 5 functional J_κ segments (Hieter *et al.* (1982) *J. Biol. Chem.*, **257**: 1516) and 4 functional J_λ segments (Vasicek and Leder (1990) *J. Exp. Med.*, **172**: 609), depending on the haplotype. The V_L segment encodes the region of the polypeptide chain which forms the first and second antigen binding loops of the V_L domain (L1 and L2), whilst the V_L and J_L segments combine to form the third antigen binding loop of the V_L domain (L3). Antibodies selected from this primary repertoire are believed to be sufficiently diverse to bind almost all antigens with at least moderate affinity. High affinity antibodies are produced by "affinity maturation" of the rearranged genes, in which point mutations are generated and selected by the immune system on the basis of improved binding.

Analysis of the structures and sequences of antibodies has shown that five of the six antigen binding loops (H1, H2, L1, L2, L3) possess a limited number of main-chain conformations or canonical structures (Chothia and Lesk (1987) *J. Mol. Biol.*, **196**: 901; Chothia *et al.* (1989) *Nature*, **342**: 877). The main-chain conformations are determined by (i) the length of the antigen binding loop, and (ii) particular residues, or types of residue, at certain key position in the antigen binding loop and the antibody framework. Analysis of the loop lengths and key residues has enabled us to the predict the main-chain conformations of H1, H2, L1, L2 and L3 encoded by the majority of human antibody sequences (Chothia *et al.* (1992) *J. Mol. Biol.*, **227**: 799; Tomlinson *et al.* (1995) *EMBO J.*, **14**: 4628; Williams *et al.* (1996) *J. Mol. Biol.*, **264**: 220). Although the H3 region is much more diverse in terms of sequence, length and structure (due to the use of D segments), it also forms a limited number of main-chain conformations for short loop lengths which depend on the length and the presence of particular residues, or types of residue, at key positions in the loop and the antibody framework (Martin *et al.* (1996) *J. Mol. Biol.*, **263**: 800; Shirai *et al.* (1996) *FEBS Letters*, **399**: 1).

A similar analysis of side-chain diversity in human antibody sequences has enabled the separation of the pattern of sequence diversity in the primary repertoire from that created by somatic hypermutation. It was found that the two patterns are complementary: diversity in the primary repertoire is focused at the centre of the antigen binding whereas somatic hypermutation spreads diversity to regions at the periphery that are highly conserved in the primary repertoire (Tomlinson *et al.* (1996) *J. Mol. Biol.*, **256**: 813; Ignatovich *et al.* (1997) *J. Mol. Biol*, **268**: 69). This complementarity seems to have evolved as an efficient strategy for searching sequence space, given the limited number B cells available for selection an any given time. Thus, antibodies are first selected from the primary repertoire based on diversity at the centre of the binding site. Somatic hypermutation is then left to optimise residues at the periphery without disrupting favourable interactions established during the primary response.

The recent advent of phage-display technology (Smith (1985) *Science*, **228**: 1315; Scott and Smith (1990) *Science*, **249**: 386; McCafferty *et al.* (1990) *Nature*, **348**: 552) has enabled the *in vitro* selection of human antibodies against a wide range of target antigens from "single pot" libraries. These phage-antibody libraries can be grouped into two categories: natural libraries which use rearranged V genes harvested from human B cells (Marks *et al.* (1991) *J. Mol. Biol.*, **222**: 581; Vaughan *et al.* (1996) *Nature Biotech.*, **14**: 309) or synthetic libraries whereby germline V gene segments are

'rearranged' *in vitro* (Hoogenboom & Winter (1992) *J. Mol. Biol.*, **227**: 381; Nissim *et al.* (1994) *EMBO J.*, **13**: 692; Griffiths *et al.* (1994) *EMBO J.*, **13**: 3245; De Kruif *et al.* (1995) *J. Mol. Biol.*, **248**: 97) or where synthetic CDRs are incorporated into a single rearranged V gene (Barbas *et al.* (1992) *Proc. Natl. Acad. Sci. USA*, **89**: 4457). Although synthetic libraries help to overcome the inherent biases of the natural repertoire which can limit the effective size of phage libraries constructed from rearranged V genes, they require the use of long degenerate PCR primers which frequently introduce base-pair deletions into the assembled V genes. This high degree of randomisation may also lead to the creation of antibodies which are unable to fold correctly and are also therefore non-functional. Furthermore, antibodies selected from these libraries may be poorly expressed and, in many cases, will contain framework mutations that may effect the antibodies immunogenicity when used in human therapy.

Recently, in an extension of the synthetic library approach it has been suggested (WO97/08320, Morphosys) that human antibody frameworks can be pre-optimised by synthesising a set of 'master genes' that have consensus framework sequences and incorporate amino acid substitutions shown to improve folding and expression. Diversity in the CDRs is then incorporated using oligonucleotides. Since it is desirable to produce artificial human antibodies which will not be recognised as foreign by the human immune system, the use of consensus frameworks which, in most cases, do not correspond to any natural framework is a disadvantage of this approach. Furthermore, since it is likely that the CDR diversity will also have an effect on folding and/or expression, it would be preferable to optimise the folding and/or expression (and remove any frame-shifts or stop codons) after the V gene has been fully assembled. To this end, it would be desirable to have a selection system which could eliminate non-functional or poorly folded/expressed members of the library before selection with the target antigen is carried out.

A further problem with the libraries of the prior art is that, because the main-chain conformation is heterogeneous, three-dimensional structural modelling is difficult because suitable high resolution crystallographic data may not be available. This is a particular problem for the H3 region, where the vast majority of antibodies derived from natural or synthetic have medium length or long loops and therefore cannot be modelled.

Summary of the Invention

According to the first aspect of the present invention, a method is provided for selecting a repertoire of polypeptides that has a first binding site for a generic ligand which is capable of binding functional members of the repertoire regardless of target ligand specificity and a second binding site for the target ligand, that involves:

a) binding the generic ligand to the first binding site and selecting the polypeptides bound to the generic ligand; and

b) binding the target ligand to the second binding site and selecting the polypeptides bound to the target ligand.

In a second aspect, the invention provides a library wherein the functional members have binding sites for both generic and target ligands.

In a third aspect, the invention provides a method for detecting, immobilising, purifying or immunoprecipitating one or more members of a repertoire of polypeptides previously selected according to the invention, comprising binding the members to the generic ligand.

In a fourth aspect, the invention provides a library comprising a repertoire of polypeptides of the immunoglobulin superfamily, wherein the members of the repertoire have a known main-chain conformation.

In a fifth aspect, the invention provides a method for selecting a polypeptide having a desired generic and/or target ligand binding site from a repertoire of polypeptides, comprising the steps of:

a) expressing a library according to the preceding aspects of the invention;

b) selecting the polypeptides by binding the generic and/or target ligand and selecting those which bind the generic and/or target ligand; and

c) optionally amplifying the selected polypeptide(s) which bind the generic and/or target ligand.

Repertoires of polypeptides are advantageously both generated and maintained in the form of a nucleic acid library. Therefore, in a sixth aspect, the invention provides a nucleic acid library encoding a repertoire of such polypeptides.

Brief Description of the Figures

Figure 1: Bar graph indicating positions in the V_H and V_K regions of the human antibody repertoire which exhibit extensive natural diversity and make antigen contacts (see Tomlinson *et al.* (1996) *J. Mol. Biol.*, **256**: 813). The H3 and the end of L3 are not shown in this representation although they are also highly diverse and make antigen contacts. Although sequence diversity in the human lambda genes has been thoroughly characterised (see Ignatovich *et al.* (1997) *J. Mol. Biol,* **268**: 69) very little data on antigen contacts currently exists for three-dimensional lambda structures.

Figure 2: Sequence of the scFv that forms the basis of a library according to the invention. There are currently two versions of the library: a "primary" library wherein 18 positions are varied and a "somatic" library wherein 12 positions are varied. The six loop regions H1, H2, H3, L1, L2 and L3 are indicated. CDR regions as defined by Kabat (Kabat *et al.* (1991). *Sequences of proteins of immunological interest*, U.S. Department of Health and Human Services) are underlined.

Figure 3: Analysis of functionality in a library according to the invention before and after selecting with the generic ligands Protein A and Protein L. Here Protein L is coated on an ELISA plate, the scFv supernatants are bound to it and detection of scFv binding is with Protein A-HRP. Therefore, only those scFv capable of binding both Protein A and Protein L give an ELISA signal.

Figure 4: Sequences of clones selected from libraries according to the invention, after panning with bovine ubiquitin, rat BIP, bovine histone, NIP-BSA, FITC-BSA, human leptin, human thyroglobulin, BSA, hen egg lysozyme, mouse IgG and human IgG. Underlines in the sequences indicate the positions which were varied in the respective libraries.

Detailed Description of the Invention

Definitions

Repertoire A repertoire is a population of diverse variants, for example nucleic acid variants which vary in nucleotide sequence or polypeptide variants which vary in amino acid sequence. A library according to the invention will encompass a repertoire of polypeptides or nucleic acids. According to the present invention, a repertoire of polypeptides is designed to possess a binding site for a generic ligand and a binding site

for a target ligand. The binding sites may overlap, or be located in the same region of the molecule, but their specificities will differ.

Generic ligand A generic ligand is a ligand that binds a substantial proportion of functional members in a given repertoire. Thus, the same generic ligand can bind many members of the repertoire regardless of their target ligand specificities (see below). In general, a functional member of the repertoire, which has the potential to possess a functional target ligand binding site, will also possess a functional generic ligand binding site. Since a functional generic ligand binding site will only be present if the repertoire member is expressed and folded correctly, binding of the generic ligand to its binding site provides a method for selecting a repertoire of polypeptides for these characteristics.

Target Ligand The target ligand is a ligand for which a specific binding member or members of the repertoire is to be identified. Where the members of the repertoire are antibody molecules, the target ligand may be an antigen and where the members of the repertoire are enzymes, the target ligand may be a substrate. Binding to the target ligand is dependent upon both the member of the repertoire being functional, as described above under *generic ligand*, and upon the precise specificity of the binding site for the target ligand.

Subset The subset is a part of the repertoire. In the terms of the present invention, it is often the case that only a subset of the repertoire is functional and therefore possesses a functional generic ligand binding site. Furthermore, it is also possible that only a fraction of the functional members of a repertoire (yet significantly more than would bind a given target ligand) will bind the generic ligand. These subsets are able to be selected according to the invention.

Library The term library refers to a mixture of heterogeneous polypeptides or nucleic acids. The library is composed of members, which have a single polypeptide or nucleic acid sequence. To this extent, *library* is synonymous with *repertoire*. Sequence differences between library members are responsible for the diversity present in the library. The library may take the form of a simple mixture of polypeptides or nucleic acids, or may be in the form organisms or cells, for example bacteria, viruses, animal or plant cells and the like, transformed with a library of nucleic acids. Preferably, each individual organism or cell contains only one member of the library. Advantageously, the nucleic acids are incorporated into expression vectors, in order to allow expression of the polypeptides encoded by the nucleic acids. In a preferred aspect, therefore, a

library may take the form of a population of host organisms, each organism containing one or more copies of an expression vector containing a single member of the library in nucleic acid form which can be expressed to produce its corresponding polypeptide member. Thus, the population of host organisms has the potential to encode a large repertoire of genetically diverse polypeptide variants.

Immunoglobulin superfamily This refers to a family of polypeptides which retain the immunoglobulin fold characteristic of immunoglobulin (antibody) molecules, which contains two β sheets and, usually, a conserved disulphide bond. Members of the immunoglobulin superfamily are involved in many aspects of cellular and non-cellular interactions *in vivo*, including widespread roles in the immune system (for example, antibodies, T-cell receptor molecules and the like), involvement in cell adhesion (for example the ICAM molecules) and intracellular signalling (for example, receptor molecules, such as the PDGF receptor). The present invention is applicable to all immunoglobulin superfamily molecules, since variation therein is achieved in similar ways. Preferably, the present invention relates to immunoglobulins (antibodies).

Main-chain conformation The main-chain conformation refers to the Cα backbone trace of a structure in three-dimensions. When individual hypervariable loops of antibodies or TCR molecules are considered the main-chain conformation is synonymous with the canonical structure. As set forth in Chothia and Lesk (1987) *J. Mol. Biol.*, **196**: 901 and Chothia *et al.* (1989) *Nature*, **342**: 877, antibodies display a limited number of canonical structures for five of their six hypervariable loops (H1, H2, L1, L2 and L3), despite considerable side-chain diversity in the loops themselves. The precise canonical structure exhibited depends on the length of the loop and the identity of certain key residues involved in its packing. The sixth loop (H3) is much more diverse in both length and sequence and therefore only exhibits canonical structures for certain short loop lengths (Martin *et al.* (1996) *J. Mol. Biol.*, **263**: 800; Shirai *et al.* (1996) *FEBS Letters*, **399**: 1). In the present invention, all six loops will preferably have canonical structures and hence the main-chain conformation for the entire antibody molecule will be known.

Antibody polypeptide Antibodies are immunoglobulins which are produced by B cells and form a central part of the host immune defence system in vertebrates. An antibody polypeptide, as used herein, is a polypeptide which either is an antibody or is a part of an antibody, modified or unmodified. Thus, the term antibody polypeptide includes a heavy chain, a light chain, a heavy chain-light chain dimer, a Fab fragment, a F(ab')2

fragment, a Dab fragment, or an Fv fragment, including a single chain Fv (scFv). Methods for the construction of such antibody molecules are well known in the art.

Preferred Embodiments of the Invention

Despite progress in the creation of "single pot" phage-antibody libraries, several problems still remain. Natural libraries (Marks *et al.* (1991) *J. Mol. Biol.*, **222**: 581; Vaughan *et al.* (1996) *Nature Biotech.*, **14**: 309) which use rearranged V genes harvested from human B cells are highly biased due to the positive and negative selection of the B cells *in vivo*. This can limit the effective size of phage libraries constructed from rearranged V genes. In addition, clones derived from natural libraries invariably contain framework mutations which may effect the antibodies immunogenicity when used in human therapy. Synthetic libraries (Hoogenboom & Winter (1992) *J. Mol. Biol.*, **227**: 381; Barbas *et al.* (1992) *Proc. Natl. Acad. Sci. USA*, **89**: 4457; Nissim *et al.* (1994) *EMBO J.*, **13**: 692; Griffiths *et al.* (1994) *EMBO J.*, **13**: 3245; De Kruif *et al.* (1995) *J. Mol. Biol.*, **248**: 97) can overcome the problem of bias but they require the use of long degenerate PCR primers which frequently introduce base-pair deletions into the assembled V genes. This high degree of randomisation may also lead to the creation of antibodies which are unable to fold correctly and are also therefore non-functional. In many cases it is likely that these non-functional members will outnumber the functional members in a library. Even if the frameworks can be pre-optimised for folding and/or expression (WO97/08320, Morphosys) by synthesising a set of 'master genes' with consensus framework sequences and by incorporating amino acid substitutions shown to improve folding and expression, there remains the problem of immunogenicity since, in most cases, the consensus sequences do not correspond to any natural framework. Furthermore, since it is likely that the CDR diversity will also have an effect of folding and/or expression, it would be preferable to optimise the folding and/or expression (and remove any frame-shifts or stop codons) after the V gene has been fully assembled.

A further problem with existing libraries is that because the main-chain conformation is heterogeneous, three-dimensional structural modelling is difficult because suitable high resolution crystallographic data may not be available. This is a particular problem for the H3 region, where the vast majority of antibodies derived from natural or synthetic have medium length or long loops and therefore cannot be modelled.

To this end, it would be desirable to have a selection system which could eliminate (or at least reduce the proportion of) non-functional or poorly folded/expressed members of

the library before selection against the target antigen is carried out. In addition, it would be advantageous to construct an antibody library wherein all the members have natural frameworks and have loops with known main-chain conformations.

The invention accordingly provides a method by which library members may be selected to remove non-functional members. This results in a marked reduction in the actual library size (and a corresponding increase in the quality of the library) without reducing the functional library size.

Libraries are produced by introducing variation into one or more molecules in order to generate members with altered specificities. In the case of antibodies, the antigen binding site is varied in order to modify or alter the binding specificity of the antibody. The introduction of variation, being at least partly random, may introduce changes which affect the overall function of the molecule. In addition, if this process is performed using the polymerase chain reaction (PCR) additional point mutations may be generated due to the relatively high mutation rate of the polymerase. Furthermore, if long oligonucleotides are used, frame-shift mutations are likely to be introduced. These variations may prevent or reduce the expression of the molecule or may modify the structure of the molecule affecting its ability to fold correctly. Either way, this is likely to severely restrict or even eliminate the ability of the molecule to bind its target. It is these variants which the invention seeks to remove from the library.

In order to remove the non-functional variants, a binding site for a generic ligand is selected, such that the ligand is only bound by functional molecules. For example, the generic ligand may be an antibody, in the form of a monoclonal antibody or a polyclonal mixture of antibodies. Preferably, these antibodies will bind an epitope on the members of the library which is of constant structure or sequence, which structure is liable to be absent or altered in non-functional members. Thus the binding site for the generic ligand could, for example, be a peptide tag or, in the case of antibodies, a superantigen binding site.

In a preferred aspect of the invention, the generic ligand is selected from the group consisting of a matrix of metallic ions, an organic compound, a protein, a peptide, a monoclonal antibody, a polyclonal antibody population, and a superantigen.

Binding of the generic ligand to its binding site and selection of library members bound to the generic ligand allows functional library members to be isolated from non-functional mutants, such as frame shift mutants, stop-codon mutants, folding mutants or

expression mutants which would be incapable of binding the target ligand. This step is referred to herein as selection of the polypeptide repertoire with the generic ligand.

Following the selection step with the generic ligand, the library may be screened in order to identify members which bind to the target ligand. Since the proportion of functional members after selection with the generic ligand is much higher, there will be an advantageous reduction in the number of non-specific "background" binders seen during selection with the target ligand. Thus, in a preferred aspect, the method according to the invention further comprises the step of binding the target ligand to its binding site and selecting the polypeptides bound to the target ligand. This step is referred to herein as selection of the polypeptide repertoire with the target ligand.

Although it is preferable to use the generic ligand to produce a pre-selected library for subsequent selection with different target ligands there may be occasions where it is advantageous to first select the library using the target ligand to produce a population of target ligand binders and than select this population using the generic ligand to isolate the subset thereof having a binding site for the generic ligand. Thus the invention provides a method for selecting the library by binding the target ligand to its binding site followed by selection by binding the generic ligand to its binding site.

In addition to removing non-functional members of a library the invention provides a method for isolating subsets of functional members on the basis of their ability to bind the generic ligand. In this case the starting library not only contains non-functional members and functional members which contain a binding site for the generic ligand but also functional members which do not contain a binding site for the generic ligand. Binding of the generic ligand and selection of bound library members removes both the non-functional members library and functional members which do not contain a binding site for the generic ligand. For example, it may be advantageous to select from a natural antibody library only those functional members which have a binding site for a given superantigen or monoclonal antibody. This could, for example, allow certain V gene families to be selected using a generic ligand which specifically binds those families, or kappa chains to be selected from lambda chains using an anti-kappa antibody.

As discussed above, in a preferred aspect of the invention, the generic ligand binds to a subset of library members, the subset being either all functional members in the library or the fraction thereof which contain a binding site for the generic ligand. In a further preferred aspect, by selecting with multiple generic ligands and/or under differential

binding conditions, which allows for binding of different library subsets, two or more subsets may be isolated from a library. These subsets may later be combined to form a single library, which includes incorporating two or more subsets into the same polypeptide chain, or using them separately. If two or more libraries are separately pre-selected using generic ligands and then combinatorially combined, a large library with a high proportion of functional members can be created. For example, where the subsets are heavy and light chains of antibody molecules, they are preferably pre-selected using generic ligands and then combined to form a heavy/light chain library, in which the heavy and light chains are either non-covalently associated or, preferably, are covalently linked (for example, by using V_H and V_L domains in an scFv format). Moreover, subsets may be combined which are obtained from different libraries. Thus, in the example given above, the heavy and light chain subsets may be isolated from separate heavy and light chain repertoires and then combined.

The members of the repertoires or libraries selected in the present invention advantageously belong to the immunoglobulin superfamily of molecules. In a preferred embodiment, the members are selected from the group consisting of antibody polypeptides or T-cell receptor polypeptides.

In a highly preferred aspect of the invention, where the library members are antibody polypeptides, the superantigens protein A and/or protein L are used as the generic ligands to select antibody repertoires, since they bind correctly folded V_H and V_L domains (which belong to certain V_H and V_L families), respectively, regardless of the sequence and structure of the binding site for the target ligand. In addition, protein A or another superantigen protein G can be used as generic ligands to select for folding and/or expression by binding the heavy chain constant domains of antibodies. The use of anti-kappa and anti-lambda antibodies for selecting light chain constant domains is also envisaged.

Moreover, the invention provides a library wherein the functional members have binding sites for both generic and target ligands.

Since it is advantageous that the generic ligand is able to bind all the functional members in a library, the library is preferably designed for the purposes of selecting with a particular generic ligand. If this is the case, the library design will be such that substantially all functional members contain a binding site for the generic ligand.

In a preferred aspect of the invention these members belong to the immunoglobulin superfamily. In a highly preferred aspect, the members are selected from the group consisting of antibody polypeptides or T-cell receptor polypeptides. In the case of an antibody library, it is advantageous to construct a library in which, for example, all functional members are able to bind the superantigens protein A and/or protein L.

As discussed above, another serious disadvantage of all antibody libraries of the prior art (in addition to the fact that they may contain a large proportion of non-functional or poorly expressed and/or folded members) is the inability to be able to predict the main-chain conformation of the vast majority of antibodies derived from these libraries.

The members of the immunoglobulin superfamily all share a similar fold for their polypeptide chain. Although antibodies and T-cell receptor molecules are, by nature, highly diverse in terms of their primary sequence, it was shown in 1987, by comparison of sequences and crystallographic structures, that contrary to expectation, five of the six antigen binding loops of antibodies (H1, H2, L1, L2, L3) adopt a limited number of main-chain conformations, or canonical structures (Chothia and Lesk (1987) *J. Mol. Biol.*, **196**: 901 and Chothia *et al.* (1989) *Nature*, **342**: 877). Analysis of the loop lengths and key residues has therefore enabled us to the predict the main-chain conformations of H1, H2, L1, L2 and L3 encoded by the majority of human antibody sequences (Chothia *et al.* (1992) *J. Mol. Biol.*, **227**: 799; Tomlinson *et al.* (1995) *EMBO J.*, **14**: 4628; Williams *et al.* (1996) *J. Mol. Biol.*, **264**: 220). Although the H3 region is much more diverse in terms of sequence, length and structure (due to the use of D segments), it also forms a limited number of main-chain conformations for short loop lengths which depend on the length and the presence of particular residues, or types of residue, at key positions in the loop and the antibody framework (Martin *et al.* (1996) *J. Mol. Biol.*, **263**: 800; Shirai *et al.* (1996) *FEBS Letters*, **399**: 1). Thus, it is possible to design a library of antibody polypeptides in which, by choosing certain loop lengths and certain key residues, the main-chain conformation of the functional members of that library is known. Since it is highly advantageous in terms of computer modelling to know the main-chain conformation of the library member, in a preferred aspect, the invention therefore provides a library wherein the functional members have a known main-chain conformation. Advantageously, this is a real conformation of an immunoglobulin superfamily molecule found in nature. It is to be understood, however, that occasional variations may occur such that a small number of functional members may possess an alternative main-chain conformation, which may be unknown.

Since a further disadvantage of libraries of the prior art is that many members have unnatural frameworks or contain framework mutations, in the case of antibodies or T-cell receptors, it is advantageous that germline V gene segments are used as a basis for constructing libraries wherein the members have a known main-chain conformation. Thus, in a highly preferred aspect, the invention provides a library wherein the repertoire of polypeptides with a known main-chain conformation is based on the use of germline V gene segment sequences as a scaffold.

In addition to being able to predict the main-chain conformation for certain sequences we can use the canonical structure theory to assess the number of different main-chain conformations encoded by human antibodies. For example, it is now known that, in the human V_K domain, the L1 loop can adopt one of four canonical structures, the L2 loop has a single canonical structure and that 90% of human V_K domains adopt one of four or five canonical structures for the L3 loop (Tomlinson *et al.* (1995) *EMBO J.*, **14**: 4628). Thus, in the V_K domain alone, different canonical structures can combine to create a significant number of main-chain conformations. Given that the $V\lambda$ domain encodes a different range of canonical structures for the L1 and L3 loops (the canonical structure for the L2 loop is the same as in the V_K domain) and that V_K and $V\lambda$ domains can pair with any V_H domain which can encode several canonical structures for the H1 and H2 loops, the number of canonical structure combinations observed for these five loops is very large. This implies that the generation of diversity in the main-chain conformation may be essential for the production of a wide range of binding specificities. However, by constructing an antibody library based on a single known main-chain conformation it was found, contrary to expectation, that diversity in the main-chain conformation is not required to generate sufficient diversity to target substantially all antigens. Even more surprisingly, the single main-chain conformation need not be a consensus structure - a single naturally occurring conformation can be used as the basis for an entire library. Thus, in a preferred aspect, the invention provides a library in which the members encode a single known main-chain conformation. It is to be understood, however, that occasional variations may occur such that a small number of functional members may possess an alternative main-chain conformation, which may be unknown.

The single main-chain conformation that is chosen is preferably commonplace amongst molecules of the immunoglobulin superfamily type in question. A conformation is commonplace when a significant number of naturally occurring molecules are observed to adopt it. Accordingly, in a preferred aspect of the invention, the natural occurrence of the different main-chain conformations for each binding loop of an immunoglobulin

superfamily molecule are considered separately and then a naturally occurring immunoglobulin superfamily molecule is chosen which possesses the desired combination of main-chain conformations for the different loops. If none is available, the nearest equivalent may be chosen. Since a disadvantage of libraries of the prior art is that many members have unnatural frameworks or contain framework mutations (see above), in the case of antibodies or T-cell receptors, it is preferable that the desired combination of main-chain conformations for the different loops is created by selecting germline gene segments which encode the desired main-chain conformations. It is more preferable, that the selected germline gene segments are frequently expressed and most preferable that they are the most frequently expressed.

In antibodies, therefore, the incidence of the different main-chain conformations for each of the six antigen binding loops can be considered separately. For H1, H2, L1, L2 and L3, preferably, between 20% and 100% of the antigen binding loops of naturally occurring molecules are observed to adopt the chosen conformation. More preferably, the incidence is above 35% (i.e. between 35% and 100%), advantageously above 50% and most preferably above 65%. In the case of H3, since the vast majority of loops do not have canonical structures, it is preferable to select a main-chain conformation which is commonplace amongst the structures which display canonical structures. For each of the loops, it is most preferable to select the conformation which is observed most often in the natural repertoire. For example, in human antibodies the most popular canonical structures (CS) for each loop are as follows: H1 - CS 1 (79% of the expressed repertoire), H2 - CS 3 (46%), L1 - CS 2 of V_K (39%), L2 - CS 1 (100%), L3 - CS 1 of V_K (36%) (calculations assumes a κ:λ ratio of 70:30, Hood *et al.* (1967) *Cold Spring Harbor Symp. Quant. Biol.*, **48**: 133). For H3 loops which have canonical structures, a CDR3 length (Kabat *et al.* (1991). *Sequences of proteins of immunological interest*, U.S. Department of Health and Human Services) of seven residues with a salt-bridge from residue 94 to residue 101 appears to be the most common. There are at least 16 human antibody sequences in the EMBL data library with the required H3 length and key residues to form this conformation and at least two crystallographic structures in the protein data bank which can be used as a basis for antibody modelling (2cgr and 1tet). In this case, the most frequently expressed germline gene segments which encode the desired combinations of main-chain conformations are the V_H segment 3-23 (DP-47), the J_H segment JH4b, the V_K segment O2/O12 (DPK9) and the J_K segment J_K1. These segments can therefore be used in combination as a basis to construct a library with the desired single main-chain conformation.

Alternatively, instead of choosing the single main-chain conformation based on the natural occurrence of the different main-chain conformations for each of the binding loops in isolation, the natural occurrence of combinations of main-chain conformations could be used as the basis for choosing the single main-chain conformation. In the case of antibodies, for example, the natural occurrence of canonical structure combinations for any two, three, four, five or for all six of the antigen binding loops could be determined. Here, it is preferable that the chosen conformation is commonplace in naturally occurring antibodies and most preferable that it observed most frequently in the natural repertoire. Thus, in human antibodies, for example, when natural combinations of the five antigen binding loops, H1, H2, L1, L2 and L3, are considered, the most frequent combination of canonical structures could be determined and then combined with the most popular conformation for the H3 loop, as a basis for choosing the single main-chain conformation.

Having selected several known main-chain conformations or, preferably a single known main-chain conformation, the library of the invention is constructed by varying the binding site of the molecule in order to generate a repertoire with structural and/or functional diversity. This means that variants are generated such that they possess sufficient diversity in their structure and/or in their function so that they are capable of providing a range of activities. For example, where the polypeptides in question are antibodies, the variants may possess a diversity of antigen binding specificities.

The desired diversity is preferably generated by varying the selected molecule at one or more positions. The positions to be changed can be chosen at random or are preferably selected. The variation can then be achieved either by randomisation, during which the resident amino acid is replaced by any amino acid or analogue thereof, natural or synthetic, producing a very large number of variants or by replacing the resident amino acid with one or more of a defined subset of amino acids, producing a more limited number of variants.

Various methods have been reported for introducing such diversity. Error prone PCR (Hawkins *et al.* (1992) *J. Mol. Biol.*, **226**: 889), chemical mutagenesis (Deng *et al.* (1994) *J. Biol. Chem.* **269**: 9533) or bacterial mutator strains (Low et al. (1996) *J. Mol. Biol.*, **260**: 359) can be used to introduce random mutations into the genes that encode the molecule. Methods for mutating selected positions are also well known in the art and include the use of mismatched oligonucleotides or degenerate oligonucleotides, with or without the use of PCR. For example, several synthetic antibody libraries have been created by targeted mutations to the antigen binding loops.

Barbas *et al.* (1992) *Proc. Natl. Acad. Sci. USA*, **89**: 4457, randomised the H3 region of a human tetanus toxoid-binding Fab to create a range of new binding specificities. Hoogenboom & Winter (1992) *J. Mol. Biol.*, **227**: 381 and later Nissim *et al.* (1994) *EMBO J.*, **13**: 692, Griffiths *et al.* (1994) *EMBO J.*, **13**: 3245 and De Kruif *et al.* (1995) *J. Mol. Biol.*, **248**: 97, appended random or semi-random H3 and L3 regions to germline V gene segments to produce large libraries with unmutated framework regions. Crameri *et al.* (1996) *Nature Med.* **2**: 100, Riechmann *et al.* (1995) *Bio/Tech.*, **13**: 475 and Morphosys (WO97/08320) extended the diversification to include some or all of the other antigen binding loops.

Since loop randomisation has the potential to create approximately more than 10^{15} structures for H3 alone and a similarly large number of variants for the other five loops, it is not feasible using current transformation technology or even by using cell free systems to produce a library representing all possible combinations. For example, in one of the largest libraries constructed to date, Griffiths *et al.* (1994) *EMBO J.*, **13**: 3245, generated $6x10^{10}$ different antibodies which is only a fraction of the potential diversity for a library of this design.

In addition to the removal of non-functional members and the use of a single known main-chain conformation, the present invention addresses these limitations by diversifying only those residues which are directly involved in creating or modifying the desired function of the molecule. For many molecules, the function will be to bind a target ligand and therefore diversity should be concentrated in the target binding site, whilst avoiding changing residues which are crucial to the overall packing of the molecule or to maintaining the chosen main-chain conformation. Thus, in a preferred aspect, the invention provides a library wherein the selected positions to be varied are those that constitute the binding site for the target ligand.

In the case of an antibody library, for example, the binding site for the target ligand is the antigen binding site. Thus, in a highly preferred aspect, the invention provides an antibody library in which only those residues in the antigen binding site are varied. These residues are highly diverse in the human antibody repertoire and are known to make contacts in high resolution antibody-antigen complexes. For example, in L2 it is known that positions 50 and 53 are diverse in naturally occurring antibodies and are observed to make contact with the antigen. In contrast, the conventional approach would have been to diversify all the residues in the corresponding Complementarity Determining Region (CDR1) as defined by Kabat *et al.* (1991). *Sequences of proteins of immunological interest*, U.S. Department of Health and Human Services, some seven

residues compared to the two diversified in the library according to the invention. This represents a significant improvement in terms of the functional diversity required to create a range of antigen binding specificities.

In nature, antibody diversity is the result of two processes: somatic recombination of germline V, D and J gene segments to create a naive primary repertoire (so called germline and junctional diversity) and somatic hypermutation of the resulting rearranged V genes. Analysis of human antibody sequences has shown that diversity in the primary repertoire is focused at the centre of the antigen binding site whereas somatic hypermutation spreads diversity to regions at the periphery of the antigen binding site that are highly conserved in the primary repertoire (see Tomlinson *et al.* (1996) *J. Mol. Biol.* **256**: 813). This complementarity has probably evolved as an efficient strategy for searching sequence space and although apparently unique to antibodies it could easily be applied to other polypeptide repertoires according to the invention. Thus, in a preferred aspect, the invention provides a library wherein the residues to be varied are a subset of those that form the binding site for the target ligand. In addition, different, including overlapping, subsets of residues in the target ligand binding site can be diversified at different stages during selection.

In the case of an antibody repertoire, a two step process may be employed, similar to that used by the human immune system. An initial naive repertoire is created where some, but not all, of the residues in the antigen binding site are diversified. This repertoire is then selected against a range of antigens. If required, further diversity can then be introduced outside the region diversified in the initial repertoire. This matured repertoire can be selected for modified function, specificity or affinity. The *in vitro* affinity maturation of antibodies using error prone PCR (Hawkins *et al.* (1992) *J. Mol. Biol.*, **226**: 889), mutator strains (Low et al. (1996) *J. Mol. Biol.*, **260**: 359), chain shuffling (Marks *et al.* (1992) *Bio/Tech.*, **10**: 779) or targeted mutagenesis (Schier *et al.* (1996) *J. Mol. Biol.*, **263**: 551) is well known in the art.

The invention provides two different naive repertoires of antibodies wherein a subset of the residues in the antigen binding site are varied. The "primary" library attempts to mimics the natural primary repertoire, with diversity restricted to residues at the centre of the antigen binding site that are diverse in the germline V gene segments (germline diversity) or diversified during the recombination process (junctional diversity). Preferably, one or more of the following residues are diversified: H50, H52, H52a, H53, H55, H56, H58, H95, H96, H97, H98, L50, L53, L91, L92, L93, L94 and L96. In the "somatic" library, diversity is restricted to residues that are diversified during the

recombination process (junctional diversity) or are highly somatically mutated. Preferably, one or more of the following residues are diversified: H31, H33, H35, H95, H96, H97, H98, L30, L31, L32, L34 and L96. In both libraries all residues are known to make contacts in one or more antibody-antigen complexes. Since in both libraries, not all the residues in the antigen binding site are varied, additional diversity can be incorporated during selection by varying the remaining residues.

In the construction of libraries according to the invention, variation of selected positions is preferably achieved at the nucleic acid level, by altering the coding sequence which specifies the sequence of the polypeptide such that a number of possible amino acids can be incorporated at that position. Thus, in a preferred aspect, the invention provides a library wherein the variation is achieved by incorporating all 20 different amino acids at the positions to be varied. Using the IUPAC nomenclature, the most versatile codon is NNK, which encodes all amino acids as well as the TAG stop codon. The NNK codon is preferably used in order to introduce the required diversity. Other codons which achieve the same ends may be used, for example, the NNN codon, but this will lead to the production of the additional stop codons TGA and TAA.

In a preferred embodiment, however, the number of different amino acids incorporated at the selected positions is more limited. Thus, in a highly preferred aspect, the invention provides a library wherein the variation is achieved by incorporating some but not all of the 20 different amino acids at the positions to be varied.

A feature of side-chain diversity in the antigen binding site of human antibodies is a pronounced bias which favours certain amino acid residues. If the amino acid composition of the ten most diverse positions in each of the V_H, V_κ and V_λ regions are summed, more than 76% of the side-chain diversity comes from only seven different residues, these being, serine (24%), tyrosine (14%), asparagine (11%), glycine (9%), alanine (7%), aspartate (6%) and threonine (6%). This bias towards hydrophilic residues and small residues which can provide main chain flexibility probably reflects the evolution of surfaces which are predisposed to binding a wide range of antigens and may help to explain the required promiscuity of antibodies in the primary repertoire.

Since it is preferable to mimic this distribution of amino acids, the invention provides a library wherein the distribution of amino acids at the positions to be varied mimics that seen in the antigen binding site of antibodies. Since the incorporation of these amino acids may be advantageous for selecting any polypeptides (not just antibody

polypeptides) against a range of target ligands this bias in amino acid variation could easily be applied to any other polypeptide repertoire according to the invention. Although there are various methods for biasing the amino acid distribution at the position to be varied (including the use of tri-nucleotide mutagenesis, WO97/08320 Morphosys), the preferred method, due to ease of synthesis, is the use of conventional degenerate codons. By comparing the amino acid profile encoded by all combinations of degenerate codons (with single, double, triple and quadruple degeneracy in equal ratios at each position) with the natural amino acid use it is possible to calculate the most representative codon. The codons (AGT)(AGC)T, (AGT)(AGC)C and (AGT)(AGC)(CT) - that is, DVT, DVC and DVY, respectively using IUPAC nomenclature - are those closest to the desired amino acid profile: they encode 22% serine and 11% tyrosine, asparagine, glycine, alanine, aspartate, threonine and cysteine. Preferably, therefore, libraries are constructed using either the DVT, DVC or DVY codon at each of the diversified positions.

The implications of all the above in terms of reduced immunogenicity, reduced library size and increased ease of library manipulation are highly advantageous. In addition, the characterisation of selected library members and their subsequent three-dimensional modelling is vastly simplified.

Preferably, the library according to the invention is an antibody polypeptide library. Antibody polypeptides may be whole antibodies or modified or unmodified fragments thereof, such as Fab, $F(ab')_2$ or Fv fragments, or separate V_H or V_L domains. Especially preferred are single chain Fv fragments, or scFvs.

ScFv fragments are reliably generated by antibody engineering methods well known in the art. The first step generally involves obtaining the genes that encode the V_H and V_L domains. These V genes may be isolated from a specific hybridoma cell line, from a B cell population, selected from a combinatorial V gene library, or made by V gene synthesis. The scFv is formed by connecting the V_H and V_L genes using an oligonucleotide that encodes an appropriately designed linker peptide, such as (Gly-Gly-Gly-Gly-Ser)3 or equivalent linker peptide(s). The linker bridges the C-terminal end of the first V region and N-terminal end of the second V region, ordered as either V_H-linker-V_L or V_L-linker-V_H. In principle, the scFv binding site can faithfully reproduce the specificity of the corresponding whole antibody.

Similar techniques are available for the construction of Fv, Fab and $F(ab')_2$ fragments, as well as chimeric antibody molecules. In essence, V_H and V_L polypeptides are

obtained as described above and expressed in the absence of a linker molecule, in order to obtain Fv fragments. When expressing Fv fragments, precautions should be taken to ensure correct chain folding and association. For Fab and F(ab')2 fragments, V_H and V_L polypeptides are combined with constant region segments, which may be isolated from rearranged genes, germline C genes or synthesised from antibody sequence data as for V region segments.

In an alternative embodiment, the library may be a V_H or V_L library. Thus, separate libraries comprising single V_H and V_L domains may be constructed and, optionally, include C_H or C_L domains, respectively, creating a Dab molecule.

Libraries according to the invention can be used for direct screening using the generic and/or target ligands or used in a selection protocol which involves the a genetic display package.

Libraries of antibodies have been generated in bacteriophage lambda expression systems which may be screened directly as bacteriophage plaques or as colonies of lysogens (Huse *et al.* (1989) *Science* **246**: 1275; Caton and Koprowski (1990) *Proc. Natl. Acad. Sci USA*, **87**: 6450; Mullinax *et al.* (1990) *Proc. Natl. Acad. Sci USA,* **87**: 8095; Persson *et al.* (1991) *Proc. Natl. Acad. Sci USA,* **88**: 2432). Such expression libraries are, however, not suited to screening of large numbers of library members (greater than 10^6 members). Other screening systems rely, for example, on direct chemical synthesis of library members. One early method involves the synthesis of peptides on a set of pins or rods, such as described in WO84/03564. A similar method involving peptide synthesis on beads, which forms a peptide library in which each bead is an individual library member, is described in U.S. Patent 4,631,211 and a related method is described in WO92/00091. A significant improvement of the bead-based methods involves tagging each bead with a unique identifier tag, such as an oligonucleotide, so as to facilitate identification of the amino acid sequence of each library member. These improved bead-based methods are described in WO93/06121.

Another chemical synthesis method involves the synthesis of arrays of peptides (or peptidomimetics) on a surface in a manner that places each distinct library member (e.g., unique peptide sequence) at a discrete, predefined location in the array. The identity of each library member is determined by its spatial location in the array. The locations in the array where binding interactions between a predetermined molecule (e.g., a receptor) and reactive library members occur is determined, thereby identifying the sequences of the reactive library members on the basis of spatial location. These

methods are described in U.S. Patent 5,143,854; WO90/15070 and WO92/10092; Fodor *et al.* (1991) *Science* **251**: 767; and Dower and Fodor (1991) *Ann. Rep. Med. Chem.* **26**: 271.

Alternatively, in a preferred aspect of the invention, a nucleic acid library encoding a repertoire of polypeptides as described in the first aspect of the invention is expressed in a selection display system, which enables the nucleic acid to be linked to the polypeptide it expresses. As used herein, a selection display system is a system which permits the selection, by suitable display means, of the individual members of the library by binding the generic and/or target ligands.

Any selection display system may be used in conjunction with a library according to the invention. Selection protocols for isolating desired members of large libraries are known in the art, as typified by phage display techniques. Systems in which diverse peptide sequences are displayed on the surface of filamentous bacteriophage (Scott and Smith (1990) *Science*, **249**: 386) have proven useful for creating libraries of antibody fragments (and the nucleotide sequences that encoding them) for the *in vitro* selection and amplification of specific antibody fragments that bind a target antigen. The nucleotide sequences encoding the V_H and V_L regions are linked to gene fragments which encode leader signals that direct them to the periplasmic space of E. coli and as a result the resultant antibody fragments are displayed on the surface of the bacteriophage, typically as fusions to bacteriophage coat proteins (e.g., pIII or pVIII). Antibody fragments can also be displayed externally on lambda phage capsids (phagebodies).

Various embodiments of bacteriophage antibody display libraries and lambda phage expression libraries have been described (McCafferty *et al.* (1990) *Nature,* **348**: 552; Kang *et al.* (1991) *Proc. Natl. Acad. Sci. USA*, **88**: 4363; Clackson *et al.* (1991) Nature, **352**: 624; Lowman *et al.* (1991) *Biochemistry*, **30**: 10832; Burton *et al.* (1991) *Proc. Natl. Acad. Sci USA,* **88**: 10134; Hoogenboom *et al.* (1991) *Nucleic Acids Res.*, **19**: 4133; Chang *et al.* (1991) *J. Immunol.*, **147**: 3610; Breitling *et al.* (1991) *Gene*, **104**: 147; Marks *et al.* (1991) *J. Mol. Biol.*, **222**: 581; Barbas *et al.* (1992) *Proc. Natl. Acad. Sci USA*, **89**: 4457; Hawkins and Winter (1992) *J. Immunol.*, **22**: 867; Marks *et al.* (1992) *Bio/Technology*, **10**: 779; Marks *et al.* (1992) *J. Biol. Chem.*, **267**: 16007; Lerner *et al.* (1992) *Science*, **258**: 1313, incorporated herein by reference).

One particularly advantageous approach has been the use of scFv phage-libraries (Huston *et al.* (1988) *Proc. Natl. Acad. Sci USA,* **85**: 5879; Chaudhary *et al.* (1990)

Proc. Natl. Acad. Sci USA, **87**: 1066; McCafferty *et al.* (1990) *Op. Cit.*; Clackson *et al.* (1991) *Op. Cit.*; Marks *et al.* (1991) *J. Mol. Biol.*, **222**: 581; Chiswell *et al.* (1992) *Trends Biotech.* **10**: 80; Marks *et al.* (1992) *Op. Cit.*). Various embodiments of scFv libraries displayed on bacteriophage coat proteins have been described. Refinements of phage display approaches are also known, for example as described in WO96/06213 and WO92/01047 (Medical Research Council *et al.*) and WO97/08320 (Morphosys), which are incorporated herein by reference.

Other systems for generating libraries of polypeptides or nucleotides involve the use of cell-free enzymatic machinery for the *in vitro* synthesis of the library members. In one method, RNA molecules are selected by alternate rounds of selection against a target ligand and PCR amplification (Tuerk and Gold (1990) Science, **249**: 505; Ellington and Szostak (1990) *Nature*, **346**: 818). A similar technique may be used to identify DNA sequences which bind a predetermined human transcription factor (Thiesen and Bach (1990) *Nucleic Acids Res.*, **18**: 3203; Beaudry and Joyce (1992) *Science*, **257**; 635; WO92/05258 and WO92/14843). In a similar way, *in vitro* translation can be used to synthesise polypeptides as a method for generating large libraries . These methods which generally comprise stabilised polysome complexes, are described further in WO88/08453, WO90/05785, WO90/07003, WO91/02076, WO91/05058, and WO92/02536. Alternative display systems which are not phage-based, such as those disclosed in WO95/22625 and WO95/11922 (Affymax) use the polysomes to display polypeptides for selection. These and all the foregoing documents also are incorporated herein by reference.

In a preferred embodiment, the method according to the invention is performed using a bacteriophage display system. An advantage of phage based display systems is that, because they are biological systems, selected library members can be amplified simply by growing the phage containing the selected library member in bacterial cells. Furthermore, since the nucleotide sequence that encode the polypeptide library member is contained on a phage or phagemid vector, sequencing, expression and subsequent genetic manipulation is relatively straightforward.

The invention accordingly provides a method for selecting a polypeptide having a desired generic and/or target ligand binding site from a repertoire of polypeptides, comprising the steps of:

a) expressing a library according to the preceding aspects of the invention in a phage-display system;

b) selecting the polypeptides by binding the generic and/or target ligand and selecting those which bind the generic and/or target ligand; and

c) optionally amplifying the selected polypeptide(s) which bind the generic and/or target ligand.

d) optionally repeating steps a) - c).

Since the invention provides a library of polypeptides which have binding sites for both generic and a target ligands the above selection method can be applied to a selection using either the generic ligand or the target ligand. Thus, the initial library can be selected as above using the generic ligand alone and then the target ligand alone, using the target ligand alone and then the generic ligand alone or using the generic and target ligands together.

Preferably, the method according to the invention further comprises the steps of subjecting the selected polypeptide(s) to additional variation (as described above) and repeating steps a) to d).

Since the generic ligand, by its very nature, is able to bind all library members selected using the generic ligand, the method according to the invention further comprises the use of the generic ligand (or some conjugate thereof) to detect, immobilise, purify or immunoprecipitate any functional member or population of members from the library (whether selected by binding the target ligand or not).

Since the invention provides a library in which the members have a known main-chain conformation the method according to the invention further comprises the production of a three-dimensional structural model of any functional member of the library (whether selected by binding the target ligand or not). Preferably, the building of such a model involves homology modelling and/or molecular replacement. A preliminary model of the main-chain conformation can be created by comparison of the polypeptide sequence to the sequence of a known three-dimensional structure, by secondary structure prediction or by screening structural libraries. Computational software may also be used to predict the secondary structure of the polypeptide. In order to predict the conformations of the side-chains at the varied positions, a side-chain rotamer library may be employed.

In general, the nucleic acid molecules and vector constructs required for the performance of the present invention are available in the art and may be constructed

and manipulated as set forth in standard laboratory manuals, such as Sambrook *et al.*, (1989) Molecular Cloning; a Laboratory Manual, Cold Spring Harbor, USA.

The manipulation of nucleic acids in the present invention is advantageously carried out in recombinant vectors. As used herein, vector refers to a discrete element that is used to introduce heterologous DNA into cells for either the expression or replication thereof. Selection and use of vectors are well within the skill of the artisan. Many vectors are available and selection of the appropriate vector depend on its intended use, for example, whether it is to be used for DNA amplification (cloning vectors) or for DNA expression (expression vectors), the size of the DNA to be inserted into the vector, and the host cell to be transformed with the vector. Each vector contains various components depending on its function, for example, whether the vector is to be used for amplification of DNA or expression of DNA and the cell type that is to be used as a host for the vector. The vector components generally include, but are not limited to, one or more of the following: an origin of replication, one or more marker genes, an enhancer element, a promoter, a transcription termination sequence and a signal sequence.

Both cloning and expression vectors generally contain nucleic acid sequences that enable the vector to replicate in one or more selected host cells. Typically in cloning vectors, this sequence is one that enables the vector to replicate independently of the host chromosomal DNA and includes origins of replication or autonomously replicating sequences. Such sequences are well known for a variety of bacteria, yeast and viruses. The origin of replication from the plasmid pBR322 is suitable for most Gram-negative bacteria, the 2μ plasmid origin is suitable for yeast, and various viral origins (e.g. SV 40, adenovirus) are useful for cloning vectors in mammalian cells. Generally, the origin of replication is not needed for mammalian expression vectors unless these are used in mammalian cells able to replicate high levels of DNA, such as COS cells.

Advantageously, a cloning or expression vector may contain a selection gene also referred to as selectable marker. This gene encodes a protein necessary for the survival or growth of transformed host cells grown in a selective culture medium. Host cells not transformed with the vector containing the selection gene will therefore not survive in the culture medium. Typical selection genes encode proteins that confer resistance to antibiotics and other toxins, e.g. ampicillin, neomycin, methotrexate or tetracycline, complement auxotrophic deficiencies, or supply critical nutrients not available in the growth media.

Since the replication of vectors according to the present invention is most conveniently performed in *E. coli*, an *E. coli* selectable marker is advantageously included, for example, the β-lactamase gene which confers resistance to the antibiotic ampicillin. These can be obtained from *E. coli* plasmids, such as pBR322 or a pUC plasmid such as pUC18 or pUC19.

Expression vectors usually contain a promoter that is recognised by the host organism and is operably linked to the coding sequence of interest. Such a promoter may be inducible or constitutive. The term "operably linked" refers to a juxtaposition wherein the components described are in a relationship permitting them to function in their intended manner. A control sequence "operably linked" to a coding sequence is ligated in such a way that expression of the coding sequence is achieved under conditions compatible with the control sequences.

Promoters suitable for use with prokaryotic hosts include, for example, the β-lactamase and lactose promoter systems, alkaline phosphatase, the tryptophan (trp) promoter system and hybrid promoters such as the tac promoter. Promoters for use in bacterial systems will also generally contain a Shine-Delgarno sequence operably linked to the coding sequence.

In the library according to the present invention, the preferred vectors are expression vectors that enables the expression of a nucleotide sequence corresponding to a polypeptide library member. Thus, selection with the generic and/or target ligands can be performed by separate propagation and expression of a single clone expressing the polypeptide library member or by use of any selection display system. As described above, the preferred selection display system is bacteriophage display. Thus, phage or phagemid vectors may be used. The preferred vectors are phagemid vectors which have an *E. Coli.* origin of replication (for double stranded replication) and also a phage origin of replication (for production of single-stranded DNA). The manipulation and expression of such vectors is well known in the art (Hoogenboom & Winter (1992) *J. Mol. Biol.*, **227**: 381; Nissim *et al.* (1994) *EMBO J.*, **13**: 692). Briefly, the vector contains a β-lactamase gene to confer selectivity on the phagemid and a lac promoter upstream of a expression cassette that consists (N to C terminal) of a pelB leader sequence (which directs the expressed polypeptide to the periplasmic space), a multiple cloning site (for cloning the nucleotide version of the library member), optionally, one or more peptide tag (for detection), optionally, one or more TAG stop codon and the phage protein pIII. Thus, using various suppressor and non-suppressor strains of *E. Coli* and with the addition of glucose, iso-propyl thio-β-D-galactoside (IPTG) or a

helper phage, such as VCS M13, the vector is able to replicate as a plasmid with no expression, produce large quantities of the polypeptide library member only or produce phage, some of which contain at least one copy of the polypeptide-pIII fusion on their surface.

Construction of vectors according to the invention employs conventional ligation techniques. Isolated vectors or DNA fragments are cleaved, tailored, and religated in the form desired to generate the required vector. If desired, analysis to confirm that the correct sequences are present in the constructed vector can be performed in a known fashion. Suitable methods for constructing expression vectors, preparing *in vitro* transcripts, introducing DNA into host cells, and performing analyses for assessing expression and function are known to those skilled in the art. Gene presence, amplification and/or expression may be measured in a sample, for example, by conventional Southern blotting (DNA analysis), Northern blotting (RNA analysis), Western blotting (protein analysis), dot blotting (DNA, RNA or protein analysis), by *in situ* hybridisation using an appropriately labelled probe, or by sequencing. Those skilled in the art will readily envisage how these methods may be modified, if desired.

The invention is further described, for the purposes of illustration only, in the following examples.

Example 1
Antibody library design

A. Main-chain conformation

For five of the six antigen binding loops of human antibodies (L1, L2, L3, H1 and H2) there are a limited number of main-chain conformations, or canonical structures ((Chothia *et al.* (1992) *J. Mol. Biol.*, **227**: 799; Tomlinson *et al.* (1995) *EMBO J.*, **14**: 4628; Williams *et al.* (1996) *J. Mol. Biol.*, **264**: 220). The most popular main-chain conformation for each of these loops is used to provide a single known main-chain conformation according to the invention. These are: H1 - CS 1 (79% of the expressed repertoire), H2 - CS 3 (46%), L1 - CS 2 of V_K (39%), L2 - CS 1 (100%), L3 - CS 1 of V_K (36%). The H3 loop forms a limited number of main-chain conformations for short loop lengths (Martin *et al.* (1996) *J. Mol. Biol.*, **263**: 800; Shirai *et al.* (1996) *FEBS Letters*, **399**: 1). Thus, where the H3 has a CDR3 length (as defined by Kabat *et al.* (1991). *Sequences of proteins of immunological interest*, U.S. Department of Health and Human Services) of seven residues and has a lysine or arginine residue at position

H94 and an aspartate residue at position H101 a salt-bridge is formed between these two residues and in most cases a single main-chain conformation is likely to be produced. There are at least 16 human antibody sequences in the EMBL data library with the required H3 length and key residues to form this conformation and at least two crystallographic structures in the protein data bank which can be used as a basis for antibody modelling (2cgr and 1tet).

In this case, the most frequently expressed germline gene segments which encode the desired loop lengths and key residues to produce the required combinations of canonical structures are the V_H segment 3-23 (DP-47), the J_H segment JH4b, the V_K segment O2/O12 (DPK9) and the J_K segment J_K1. These segments can therefore be used in combination as a basis to construct a library with the desired single main-chain conformation. The V_K segment O2/O12 (DPK9) is member of the V_K1 family and therefore will bind the superantigen Protein L. The V_H segment 3-23 (DP-47) is a member of the V_H3 family and therefore should bind the superantigen Protein A.

B. Selection of positions for variation

Analysis of human V_H and V_K sequences indicates that the most diverse positions in the mature repertoire are those that make the most contacts with antigens (see Tomlinson *et al.*, (1996) *J. Mol. Biol.*, **256**: 813; Figure 1). These positions form the functional antigen binding site and are therefore selected for side-chain diversification (Figure 2). H54 is a key residue and points away from the antigen binding site in the chosen H2 canonical structure 3 (the diversity seen at this position is due to canonical structures 1, 2 and 4 where H54 points into the binding site). In this case H55 (which points into the binding site) is diversified instead. The diversity at these positions is created either by germline or junctional diversity in the primary repertoire or by somatic hypermutation (Tomlinson *et al.*, (1996) *J. Mol. Biol.*, **256**: 813; Figure 1). Two different subsets of residues in the antigen binding site were therefore varied to create two different library formats. In the "primary" library the residues selected for variation are from H2, H3, L2 and L3 (diversity in these loops is mainly the result of germline or junctional diversity). The positions varied in this library are: H50, H52, H52a, H53, H55, H56, H58, H95, H96, H97, H98, L50, L53, L91, L92, L93, L94 and L96 (18 residues in total, Figure 2). In the "somatic" library the residues selected for variation are from H1, H3, L1 and the end of L3 (diversity here is mainly the result of somatic hypermutation or junctional diversity). The positions varied in this library are: H31, H33, H35, H95, H96, H97, H98, L30, L31, L32, L34 and L96 (12 residues in total, Figure 2).

C. Selection of amino acid use at the positions to be varied

Side-chain diversity is introduced into the "primary" and "somatic" libraries by incorporating either the codon NNK (which encodes all 20 amino acids, including the TAG stop codon, but not the TGA and TAA stop codons) or the codon DVT (which encodes 22% serine and 11% tyrosine, asparagine, glycine, alanine, aspartate, threonine and cysteine and using single, double, triple and quadruple degeneracy in equal ratios at each position, most closely mimics the distribution of amino acid residues for in the antigen binding sites of natural human antibodies).

Example 2
Library construction and selection with the generic ligands

The "primary" and "somatic" libraries were assembled by PCR using the oligonucleotides listed in Table 1 and the germline V gene segments DPK9 (Cox *et al.* (1994) *Eur. J. Immunol.*, **24**: 827) and DP-47 (Tomlinson *et al.* (1992) *J. Mol. Biol.*, **227**: 7768). Briefly, first round of amplification was performed using pairs of 5' (back) primers in conjunction with NNK or DVT 3' (forward) primers together with the corresponding germline V gene segment as template (see Table 1). This produces eight separate DNA fragments for each of the NNK and DVT libraries. A second round of amplification was then performed using the 5' (back) primers and the 3' (forward) primers shown in Table 1 together with two of the purified fragments from the first round of amplification. This produces four separate fragments for each of the NNK and DVT libraries (a "primary" V_H fragment, 5A; a "primary" V_K fragment, 6A; a "somatic" V_H fragment, 5B; and a "somatic" V_K fragment, 6B).

Each of these fragments was cut and then ligated into pCLEANVH (for the V_H fragments) or pCLEANVK (for the V_K fragments) which contain dummy V_H and V_K domains, respectively in a version of pHEN1 which does not contain any TAG codons or peptide tags (Hoogenboom & Winter (1992) *J. Mol. Biol.*, **227**: 381). The ligations were then electroporated into the non-suppressor *E. Coli.* strain HB2151. Phage from each of these libraries was produced and separately selected using immunotubes coated with 10 μg/ml of the generic ligands Protein A and Protein L for the V_H and V_K libraries, respectively. DNA from *E. Coli.* infected with selected phage was then prepared and cut so that the dummy V_K inserts were replaced by the corresponding V_K libraries. Electroporation of these libraries results in the following insert library sizes: 9.21×10^8 ("primary" NNK), 5.57×10^8 ("primary" DVT), 1.00×10^9 ("somatic"

NNK) and 2.38 x 10^8 ("somatic" DVT). As a control for pre-selection four additional libraries were created but without selection with the generic ligands Protein A and Protein L: insert library sizes for these libraries were 1.29 x 10^9 ("primary" NNK), 2.40 x 10^8 ("primary" DVT), 1.16 x 10^9 ("somatic" NNK) and 2.17 x 10^8 ("somatic" DVT).

To verify the success of the pre-selection step, DNA from the selected and unselected "primary" NNK libraries was cloned into a pUC based expression vector and electroporated into HB2151. 96 clones were picked at random from each recloned library and induced for expression of soluble scFv fragments. Production of functional scFv is assayed by ELISA using Protein L to capture the scFv and then Protein A-HRP conjugate to detect binding. Only scFv which express functional V_H and V_K domains (no frame-shifts, stop codons, folding or expression mutations) will give a signal using this assay. The number of functional antibodies in each library (ELISA signals above background) was 5% with the unselected "primary" NNK library and 75% with the selected version of the same (Figure 3). Sequencing of clones which were negative in the assay confirmed the presence of frame-shifts, stop codons, PCR mutations at critical framework residues and amino acids in the antigen binding site which must prevent folding and/or expression.

Example 3
Library selection against target ligands

The "primary" and "somatic" NNK libraries (without pre-selection) were separately selected using five antigens (bovine ubiquitin, rat BIP, bovine histone, NIP-BSA and hen egg lysozyme) coated on immunotubes at various concentrations. After 2-4 rounds of selection, highly specific antibodies were obtained to all antigens except hen egg lysozyme. Clones were selected at random for sequencing demonstrating a range of antibodies to each antigen (Figure 4).

In the second phase, phage from the pre-selected NNK and DVT libraries were mixed 1:1 to create a single "primary" library and a single "somatic" library. These libraries were then separately selected using seven antigens (FITC-BSA, human leptin, human thyroglobulin, BSA, hen egg lysozyme, mouse IgG and human IgG) coated on immunotubes at various concentrations. After 2-4 rounds of selection, highly specific antibodies were obtained to all the antigens, including hen egg lysozyme which failed to produce positives in the previous phase of selection using the libraries that had not been

pre-selected using the generic ligands. Clones were selected at random for sequencing, demonstrating a range of different antibodies to each antigen (Figure 4).

Claims

1. A method for selecting a repertoire of polypeptides that has a first binding site for a generic ligand which is capable of binding functional members of the repertoire regardless of target ligand specificity and a second binding site for the target ligand, that involves:

a) binding the generic ligand to the first binding site and selecting the polypeptides bound to the generic ligand; and
b) binding the target ligand to the second binding site and selecting the polypeptides bound to the target ligand.

2. A method according to claim 1 wherein the repertoire of polypeptides is first selected by binding the target ligand to the second binding site and then by binding the generic ligand to the first binding site.

3. A method according to claim 1 wherein the generic ligand binds a subset of the repertoire of polypeptides.

4. A method according to claim 3 wherein the repertoire of polypeptides is selected to isolate two or more subsets thereof.

5. A method according to claim 4 wherein the selection is performed with two or more generic ligands.

6. A method according to claims 4 or 5 wherein two or more subsets are combined.

7. A method according to any preceding claim wherein two or more repertoires of polypeptides are selected with generic ligands and then combined.

8. A method according to any preceding claim, wherein the polypeptides of the repertoire are of the immunoglobulin superfamily.

9. A method according to claim 8, wherein the polypeptides are antibody or T-cell receptor polypeptides.

10. A method according to claim 9, wherein the polypeptides are V_H or $V\beta$ domains.

11. A method according to claim 9, wherein the polypeptides are V_L or $V\alpha$ domains.

12. A method wherein a repertoire of polypeptides according to claim 10 and a repertoire of polypeptides according to claim 11 are selected with generic ligands and then combined.

13. A method according to any preceding claim wherein the generic ligand is selected from the group consisting of a matrix of metallic ions, an organic compound, a protein, a peptide, a monoclonal antibody, a polyclonal antibody population, and a superantigen.

14. A method for detecting, immobilising, purifying or immunoprecipitating one or more members of a repertoire of polypeptides previously selected according to any one of claims 1 to 13, comprising binding the members to the generic ligand.

15. A library wherein the functional members have binding sites for both generic and target ligands.

16. A library designed for selection with both generic and target ligands.

17. A library according to claim 15 and 16 comprising a repertoire of polypeptides of the immunoglobulin superfamily.

18. A library according to claim 17 wherein the polypeptides are antibody or T-cell receptor polypeptides.

19. A library according to claim 18, wherein the polypeptides are V_H or $V\beta$ domains.

20. A library according to claim 18, wherein the polypeptides are V_L or $V\alpha$ domains.

21. A library wherein a repertoire of polypeptides according to claim 19 and a repertoire of polypeptides according to claim 20 are selected with generic ligands and then combined.

22. A library according to any one of claims 15 to 21, wherein the functional members of the repertoire have a known main-chain conformation.

23. A library according to claim 22, wherein the functional members of the repertoire have a single main-chain conformation.

24. A library according to claims 22 or 23, wherein the immunoglobulin scaffold is based on germline V gene segment sequences.

25. A library according to any one of claims 15 to 24, wherein the polypeptides are varied at random positions.

26. A library according to any one of claims 15 to 24, wherein the polypeptides are varied at selected positions.

27. A library according to claim 26, wherein the selected positions are those which form the binding site for the target ligand.

28. A library according to claim 27, wherein the selected positions are a subset those which form the binding site for the target ligand.

29. A library wherein a repertoire of polypeptides according to claim 28 is selected by binding the target ligand and then varied at a different subset of residues in order to modify the function, specificity or affinity of target ligand interaction.

30. A library according to claims 26-29, wherein the variation is achieved by incorporating all 20 different amino acids at the positions to be varied.

31. A library according to claim 26-29, wherein the variation is achieved by incorporating some but not all of the 20 different amino acids at the positions to be varied

32. A nucleic acid library encoding a repertoire of polypeptides according to any one of claims 15 to 31.

Abstract

The invention provides a method for selecting a repertoire of polypeptides that has a first binding site for a generic ligand which is capable of binding functional members of the repertoire regardless of target ligand specificity and a second binding site for the target ligand, that involves:

a) binding the generic ligand to the first binding site and selecting the polypeptides bound to the generic ligand; and

b) binding the target ligand to the second binding site and selecting the polypeptides bound to the target ligand.